Supplementary Notes

Relationship Between Slots, Resources, and SD

The slots+averaging model supposes that observers may store multiple independent samples of a colour in separate slots, reporting the average of the stored colour values at the time of the memory probe. The SD of the average of N samples is well known to be equal to the SD of the individual samples divided by sqrt(N)[1]. This fact was used to make quantitative predictions from the slots+averaging model.

Resource models are analogous to slot models with an infinite number of slots. Consequently, quantitative resource models also lead to a square-root relationship between the quantity of resources devoted to a representation and the SD of that representation[1,2]. The essential difference between slot and resource models is the granularity of the resource, which is infinitely divisible in resource models but is organized into a few large chunks in slot models.

Adequacy of Parameter Estimations

As evidence that our mixture model with three parameters provides an adequate description of the data, we computed the adjusted r^2 statistic (which reflects the proportion of variance explained by the model) and the χ^2 statistic on the basis of histograms of the data with 15 bins, each 24° wide. The Kolmogorov-Smirnov (K-S) statistic was also computed to test whether the observed data differ significantly from the model. These statistics were computed for each individual subject and also for the data aggregated over subjects. It is more difficult to obtain a good fit when P_m is low and when a given condition has relatively few trials per subject, and the group data sets were particularly useful for demonstrating goodness of fit under these conditions. We also conducted 1000 Monte Carlo simulations of the data that would be expected for

individual observers in each condition to determine how the adjusted r^2 values would be expected to vary given the number of trials and the observed P_m values.

Supplementary Table 1 shows the adjusted r^2 values for each major experiment described in the main text, including the values computed from the group data, the mean of the single-subject data, and the mean of the simulated single-subject data. In the single-subject data, the worst fits explained an average of over 75% of the variance, and the best fits explained an average of over 95% of the variance. The variations in goodness of fit corresponded well with variations observed in the simulations, indicating that the variations in goodness of fit are a consequence of variations in the number of trials and P_m levels rather than reflecting systematic deviations of the data from the model. For example, although the observed adjusted r^2 values fell from approximately 0.95 at set sizes 1–3 to approximately 0.87 at set size 6, a similar drop was observed in the simulations, presumably because of an increase in the proportion of trials on which subjects did not recall the probed colour and therefore responded randomly. Thus, the reduction in goodness of fit was an inevitable consequence of the nature of the underlying memory representations in the context of a finite data set. In addition, the model explained over 95% of the variance in all conditions when the data were aggregated across the group of subjects. Moreover, the K-S and χ^2 analyses indicated that the observed data were not significantly different from the model in any subject or group of subjects for any condition. Thus, this simple 3-parameter model provides an excellent quantitative fit to the data across all conditions of all experiments.

For Experiments 1 and 2, we also tested the adequacy of a simple resource model containing only a von Mises distribution, which is equivalent to holding P_m constant at 1.0 when estimating the μ and SD parameters. The adjusted r^2 values were negative for all set sizes in both experiments, with the exception of set size 1 in Experiment 2. A negative value means that the error variance is larger than what would be obtained by a

model in which the mean of the sample was the only parameter, and it indicates that the model does not adequately fit the data. In addition, the data deviated significantly from the model according to the K-S and χ^2 tests for all set sizes in both experiments ($p < .05$ or better). Thus, the data are quantitatively inconsistent with a model in which all items are encoded and only the precision of the representations varies as a function of set size[3].

Control Experiments

To demonstrate that the effect of set size on P_m did not reflect a lack of sufficient time to encode the items at set size 6, we conducted a control experiment comparing sample durations of 100 and 500 ms at set size 6. We found no significant effect of duration on either SD or P_e (both $Fs < 1$). To demonstrate that our methods are sufficiently sensitive to detect changes in SD if they are present, we conducted an additional control experiment in which the quality of the perceptual representations was manipulated by adding varying numbers of coloured "noise" dots to a set of 3 coloured squares (see Supplementary Figure 1). The noise degraded the perceptual representation of the colours, and this reduced precision was necessarily propagated to the memory representations. Increasing the noise increased the SD ($F(1,7)= 7.78$, $p<0.03$) but did not influence P_m ($F < 1$). Thus, our methods are sufficiently sensitive to detect modest changes in precision.

Colour Categories

Although it is not central to this study, our methods implicitly assume that observers store a representation of the continuous colour values. However, it is possible that they instead convert the continuous colour values in the sample array into categorical representations (e.g., prototypes of red, green, blue, etc.). If this were true, then much of the distribution of responses would reflect the difference between the

actual colour of the probed item and the nearest prototypical colour value. To assess this possibility, we pooled the data from four experiments that included trials with a set size of one item (the procedure for these trials was identical across experiments), yielding a total of 50 observers. (Pooling across observers is well justified because colour categories are highly consistent across individuals from a restricted age range and cultural group[4].) We then plotted the distribution of reported colours as a function of the actual colours for this pool of observers (Supplementary Figure 2a). If observers represent the actual colour (plus noise), then this should yield a straight line. If observers instead represent a given sample colour as the nearest colour category value, then this function should look like a staircase, in which variations in the actual colour within a given range lead to no change in the reported colour, with a sudden change in reported colour when the actual colour crosses the category boundary. An example of this is displayed in Supplementary Figure 2b, which shows the results of a simulation of categorical memory with 7 colour categories. The results shown in Supplementary Figure 2a clearly follow a straight line, and a least-squares analysis showed that a straight line accounts for 97% of the variance in reported colour. There was no sign of staircase-like horizontal bands in these data. Thus, observers appear to remember the actual colour rather than the nearest colour prototype.

Comparison of Colour Recall with Colour Change Detection

The colour recall task used in this study differs from the more common change detection task that has been used widely to study visual working memory in behavioural[5], ERP[6], and neuroimaging[7] studies. To determine whether these two tasks are measuring the same aspects of working memory, we conducted an experiment (N=14) in which the two tasks were randomly intermixed. Each trial began with a 100-ms presentation of a sample array containing three items. This was followed after a 900-ms delay by either a colour recall test display or a change-detection test display.

The colour recall test display contained a probe and a colour wheel, just as in the other experiments in the present study, and observers responded by selecting a colour from the colour wheel. The change detection test display contained a single coloured square at one of the locations from the sample array, and observers responded by making a keypress response to indicate whether or not the test square was the same colour as the corresponding sample square. The test colour was the same on 50% of trials and differed by 180° in colour space on the other 50%. The colour recall and change detection trials were unpredictably intermixed, so observers necessarily encoded and maintained the colour information in the same way on both trial types.

The goal of this experiment was to determine whether the estimated number of items that observers store in memory (the storage capacity) is the same for the two tasks. The storage capacity for the colour recall task (K_i) was estimated by simply multiplying P_m by the set size. The storage capacity for the change detection task was estimated with the Cowan K equation[8]. This equation is based on a high-threshold model, but it should be approximately correct for the maximally large (180°) change magnitudes used in this experiment. As shown in Supplementary Figure 3, the two measures of storage capacity were strongly and significantly correlated across subjects ($r^2 = .572$, $p = .002$), with a slope near 1.0 and an intercept near 0.0. Although the agreement between these two procedures may vary depending on the decision requirements of the specific experiment, these results suggest that they are measuring fundamentally similar aspects of visual working memory capacity.

Extension to Shape-Defined Stimuli

Experiments 2 and 3 were repeated with shape stimuli (N = 8 and 14, respectively; see stimuli in Supplementary Figure 4 and results in Supplementary Figure 5). Shape was parameterized using the Fourier descriptor technique[9], in which the

perimeter of an object is described by the sum of a set of sinusoidal components. That is, a function is created that represents the distance between the centre of the object and its perimeter as a function of polar angle, and this function is then decomposed into the sum of a set of sine wave components that vary in frequency, amplitude, and phase. Using this approach, we synthesized a family of objects consisting of two sinusoidal components, one with a frequency of 2 cycles per perimeter (cpp) and an amplitude of 0.5 and one with a frequency of 4 cpp and an amplitude of 0.5. The phase of the 2-cpp component was held constant at 0°, and the phase of the 4-cpp component varied between 0 and 360° in steps of 2° (providing a circular dimension that is analogous to the hue dimension used in the colour experiments). The result was a family of 180 moderately complex shapes, each subtending approximately 2° in visual angle, that varied systematically in shape.

The procedure for the shape stimuli was identical to that used for the colour stimuli, with three exceptions. First, the exposure duration was increased to 1000 ms to provide sufficient encoding time. Second, the circle of shapes at the time of response contained 30 discrete shapes rather than a visually continuous ring of 180 colour values (see Supplementary Figure 4). These shapes were sampled from the family of 180 shapes, with 12° of phase difference between adjacent shapes and a randomly chosen starting phase on each trial. Subjects were instructed to indicate the remembered shape of the probed item by clicking on the corresponding position within the circle of shapes, interpolating between the exemplars in the circle if necessary to accurately report the shape (because the actual shape may lie between two of the sample shapes in a given display). Third, a familiarization block with 60 trials was run before the memory task, in which six shapes were presented simultaneously with the circle of shapes. One of the six shapes was cued and subjects matched the cued shape to the corresponding shape on the circle of shapes. This provided the observers with an opportunity to practice choosing interpolated locations along the circle of test shapes, and the experimenter

verified that they did so. The data from the shape experiments were analyzed in exactly the same manner as the data from the corresponding colour experiments.

Overall P_m values for shape were similar to those obtained for colour. The SD was somewhat larger for shape than for colour, but there is no reason why the precision should have the same numeric value across dimensions, especially given that the test display contained 30 values for shape and 180 values for colour. When the set size was manipulated as in Experiment 2, P_m fell slightly as the set size increased from 1 to 3 and then fell dramatically as the set size increased from 3 to 6, producing a significant overall effect of set size ($F(3,21) = 51.76$, $p < 0.001$). SD increased significantly as the set size increased between 1 and 3 items ($F(2,14) = 13.15$, $p < 0.001$) but did not increase as the set size increased from 3 to 6 items ($F < 1$). This is the same pattern of results obtained for colour in Experiment 2, and it was well fit by both the slots+resources and slots+averaging models but not by the simple resource model.

When a cue in the initial sample array was used to direct attention to a single item, as in Experiment 3, P_m was high for valid trials, substantially lower on neutral trials, and very low on invalid trials ($F(2,26)=15.23$, $p<0.001$). This demonstrates that the cue was highly effective in motivating the observers to give priority to the cued shape. However, cuing produced only a modest difference in SD between valid and neutral trials ($F(1,13)=12.54$, $p<0.01$) and no difference between neutral and invalid trials ($F < 1$). This pattern is identical to the pattern observed for colour in Experiment 3. The size of the improvement in SD on valid trials compared to neutral trials was within the range that would be expected if the observers allocated multiple slots to the cued item on valid trials and allocated a single slot to each item on neutral trials. Most importantly, the lack of an increase in SD on invalid trials compared to neutral trials indicates that focusing attention onto the cued item did not result in reduced precision for the uncued items. That is, uncued items were relatively unlikely to be stored in memory, but when

they were stored they were represented with the same precision as on neutral trials. Thus, it does not seem possible to allocate “only a few drops” of resources to a shape representation in working memory.

Supplementary References

1 J. Palmer, *J. Exp. Psychol. Hum. Percept. Perform.* **16**, 332 (1990).

2 M.L. Shaw, in *Attention and performance*, edited by R. S. Nickerson (L. Erlbaum Associates Hillsdale, NJ, 1980), Vol. VIII, pp. 277; A. M. Bonnel and J. Miller, *Percept. Psychophys.* **55** (2), 162 (1994).

3 P. Wilken and W. J. Ma, *Journal of Vision* **4**, 1120 (2004).

4 E. Rosch, *Cognit. Psychol.* **7**, 532 (1975).

5 S.J. Luck and E.K. Vogel, *Nature* **390**, 279 (1997); M. Wheeler and A. M. Treisman, *J. Exp. Psychol. Gen.* **131**, 48 (2002); Y. Jiang, I.R. Olson, and M.M. Chun, *Journal of Experimental Psychology: Learning, Memory & Cognition* **2**, 683 (2000).

6 E.K. Vogel and M.G. Machizawa, *Nature* **428**, 748 (2004); E.K. Vogel, A.W. McCollough, and M.G. Machizawa, *Nature* **438**, 500 (2005).

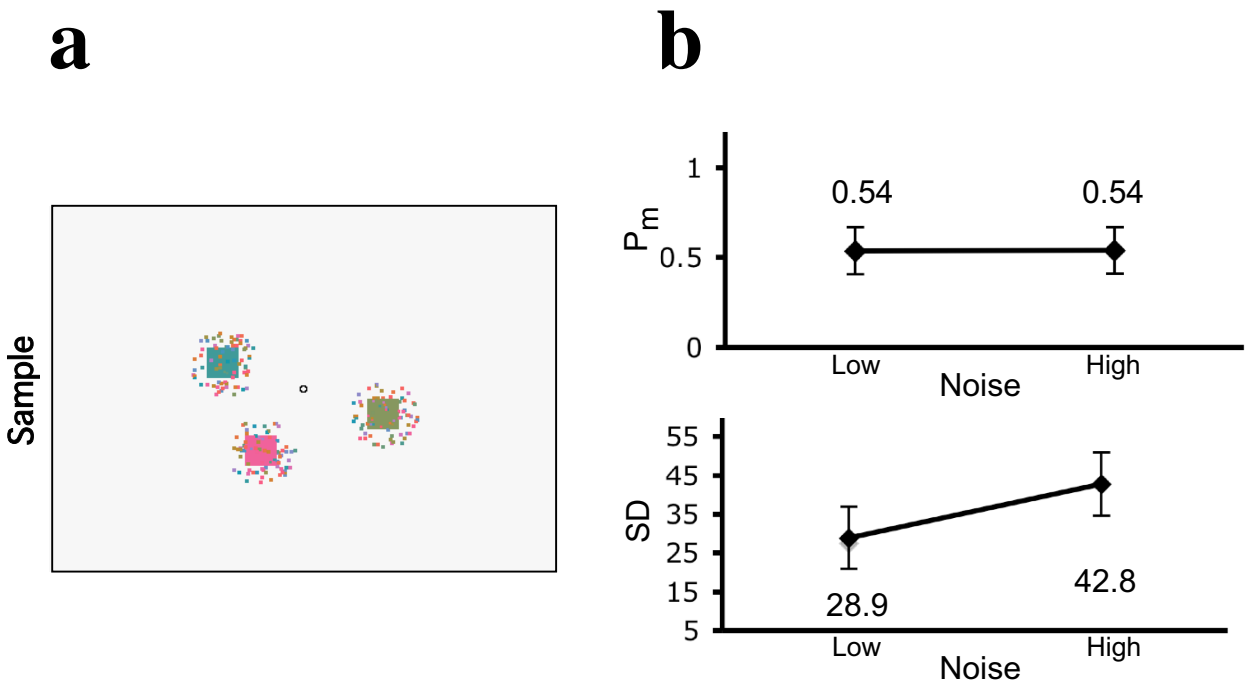
7 Y. Xu and M.M. Chun, *Nature* (2006); J. J. Todd and R. Marois, *Nature* **428**, 751 (2004).

8 N. Cowan, E.M. Elliott, J.S. Saults et al., *Cognit. Psychol.* **51**, 42 (2005).

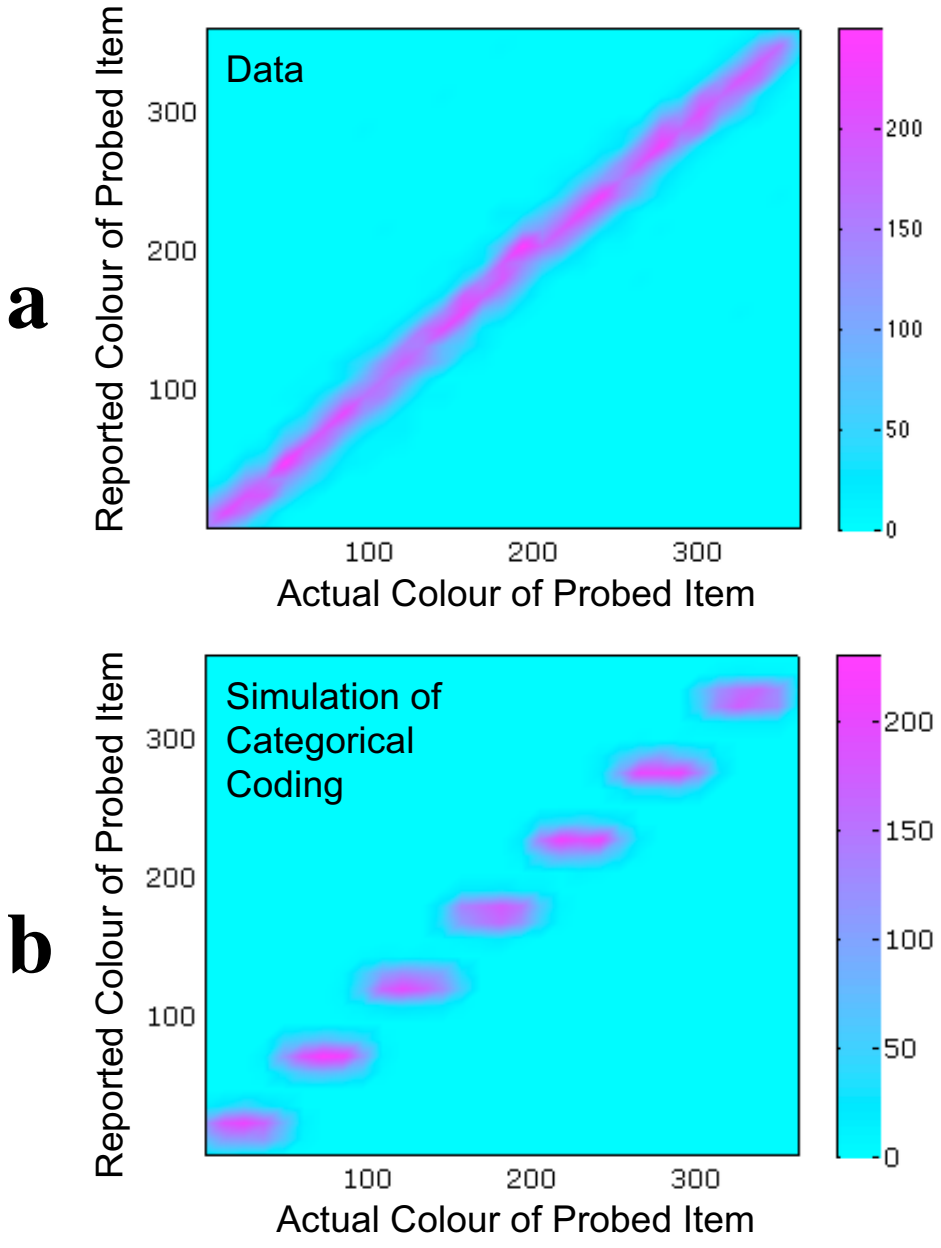
9 C. T. Zahn and R. Z. Roskies, *IEEE Transactions on Computers* **C21** (3), 269 (1972).

Supplementary Table 1. Adjusted r^2 values indicating the goodness of fit for the slots+averaging model when applied to the group data, the mean of the individual data, and the simulated individual data.

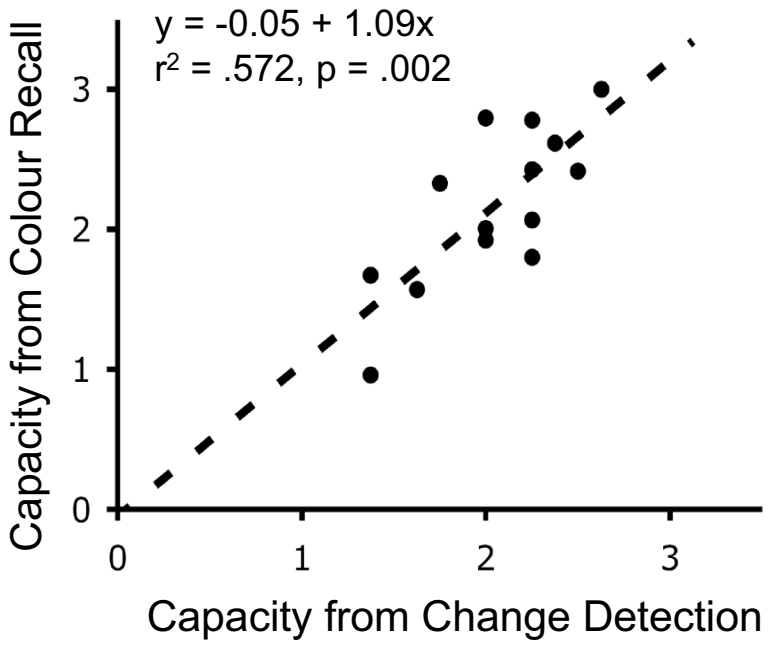
	Condition	Group r^2	Mean Individual r^2	Simulated Individual r^2
Experiment 1	Set Size 3	0.992	0.965	0.976
	Set Size 6	0.980	0.847	0.870
Experiment 2	Set Size 1	0.979	0.973	0.965
	Set Size 2	0.995	0.974	0.983
	Set Size 3	0.995	0.962	0.975
	Set Size 6	0.959	0.822	0.866
Experiment 3	Valid	0.993	0.963	0.972
	Neutral	0.986	0.835	0.906
	Invalid	0.964	0.767	0.596
Experiment 4	110-ms SOA	0.956	0.773	0.861
	340-ms SOA	0.983	0.881	0.949



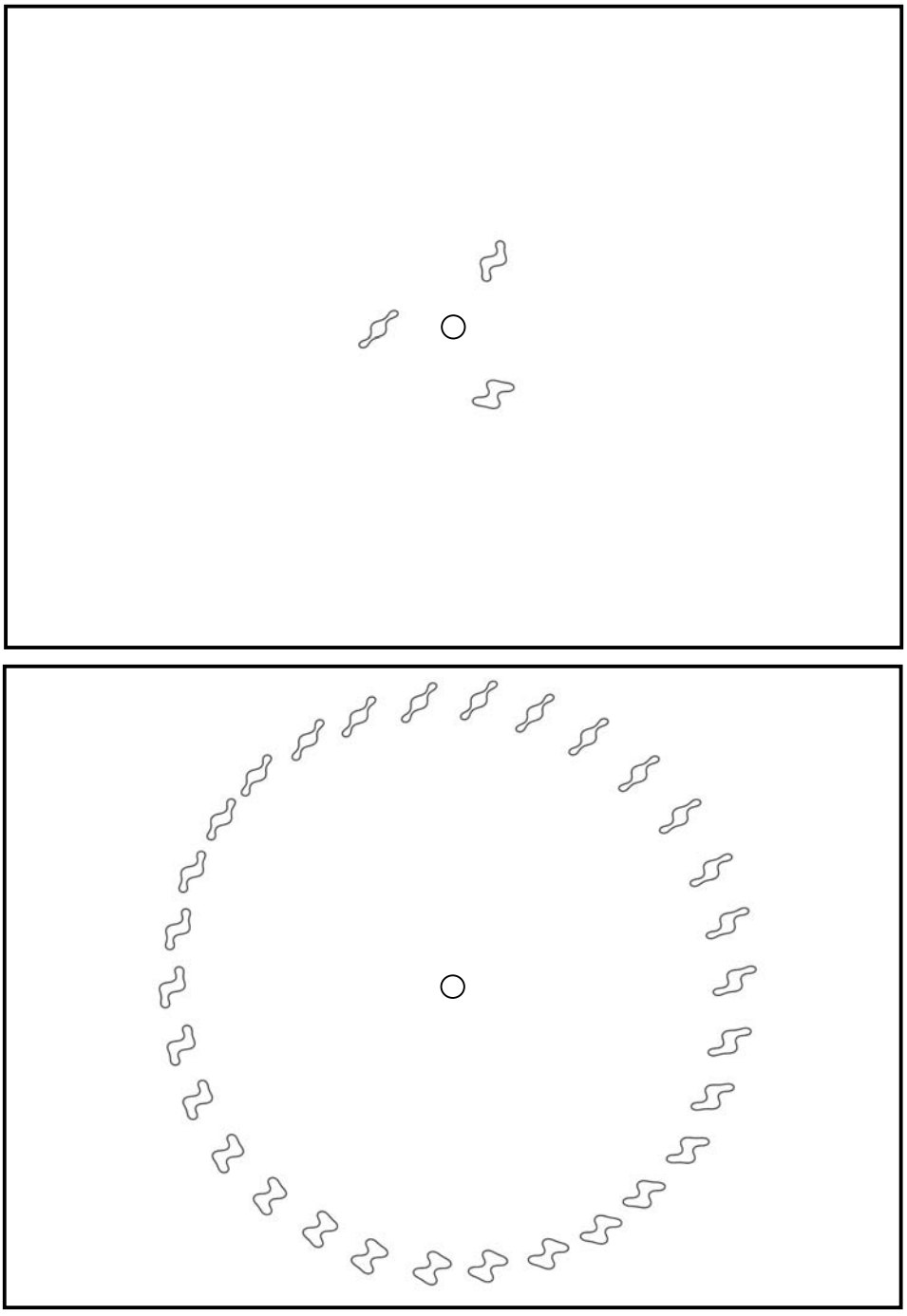
Supplementary Figure 1. Stimuli (a) and results (b) from a control experiment showing that adding sensory noise to the sample array increases the estimated standard deviation (SD) but not the estimated probability of memory (Pm). Error bars show within-subjects 95% confidence intervals.



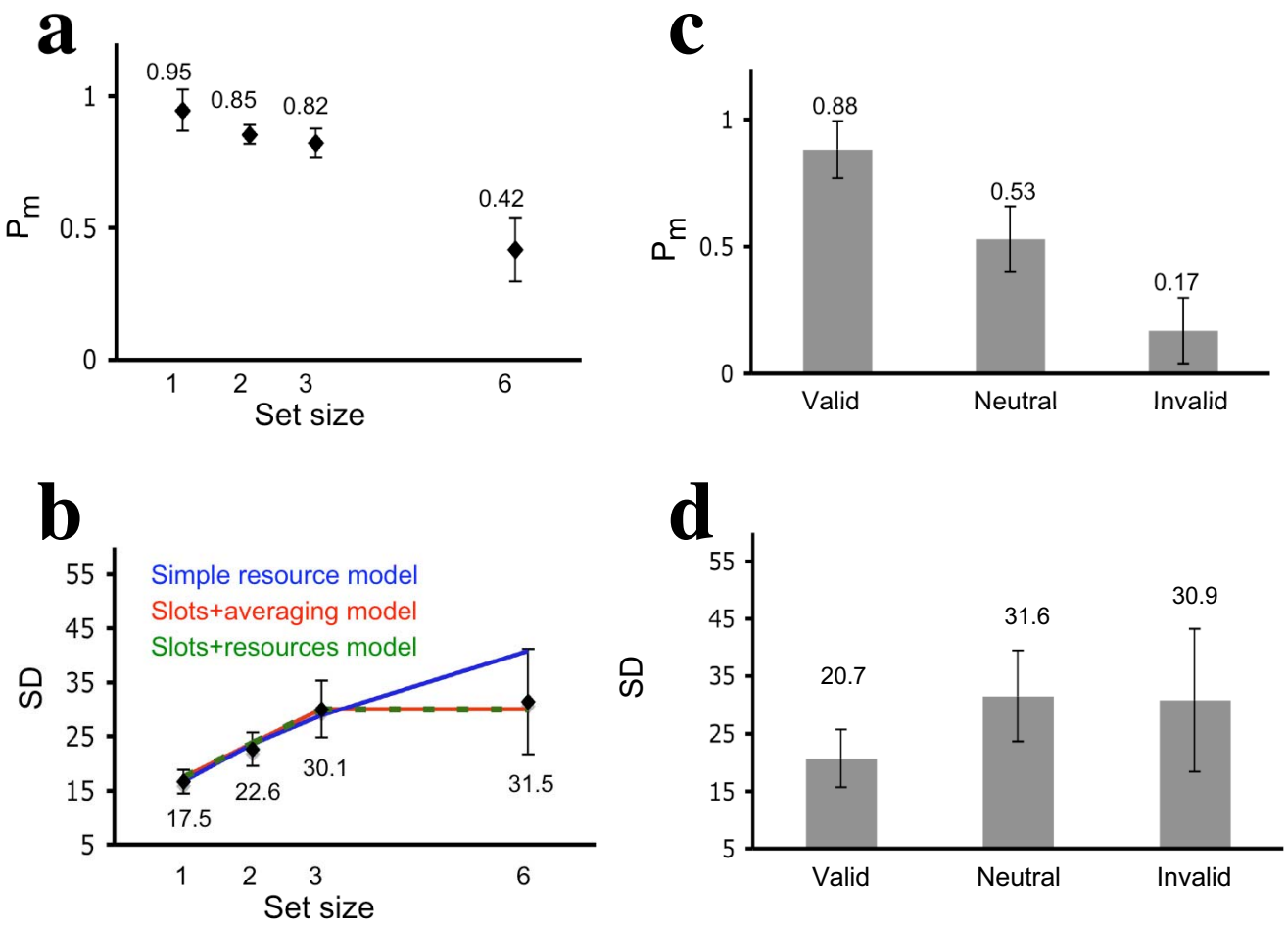
Supplementary Figure 2. **a**, Frequency of reporting a given colour value as a function of the actual colour value, aggregated across at least 150 trials for each of 50 observers across four separate experiments at set size 1. The intensity at a given point in the plot represents the frequency of occurrence for that particular combination of actual colour and reported colour. The function is nearly perfectly linear, indicating that the observers stored a representation of the actual colour (plus noise) rather than storing the nearest colour category. **b**, Analogous results from a Monte Carlo simulation of a memory system in which the actual colour value was stored as the nearest of seven equally spaced colour categories. On each trial of the simulation, Gaussian noise was added to the actual colour value, the nearest categorical value was chosen, and then the categorical colour value was reported (plus additional Gaussian noise to represent response variability). The noise levels were chosen to produce results that matched the overall level of response error exhibited by the observers in **a**. Additional simulations showed that a categorical model could not achieve the low level of response error and the linearity of the data shown by the observers in **a** unless the number of colour categories was unrealistically large (~20).



Supplementary Figure 3. Memory capacity as estimated from the colour recall task (K_i) as a function of memory capacity as estimated from a change detection task (Cowan's K).



Supplementary Figure 4. Example sample array and circle of shapes from the shape experiments. The contour of each shape can be described by the sum of two sine waves. Note that the shapes in the circle were evenly spaced in phase space, with a starting phase that varied randomly from trial to trial. The actual sample shapes were not necessarily among the exemplars shown in the circle, and the observers were instructed to make interpolated responses when the remembered shape fell between two of the exemplars in the circle. The observers were given extensive practice with making interpolated responses during a familiarization phase, in which the central sample shapes were presented simultaneously with the circle of test shapes.



Supplementary Figure 5. **(a)** P_m and **(b)** SD results from an experiment with shape stimuli in which the set size varied between values of 1, 2, 3, and 6. The lines in **b** show the predictions of the simple resource model, the slots+averaging model, and the slots+resources model. **(c)** P_m and **(d)** SD results from an experiment with shape stimuli in which a cue was presented in the sample array to indicate which item was most likely to be tested. Error bars show within-subjects 95% confidence intervals.